4.2.2 DENDRAL PROJECT

RESOURCE RELATED RESEARCH -- COMPUTERS AND CHEMISTRY: THE DENDRAL PROJECT

Carl Djerassi, Principal Investigator
Professor of Chemistry
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

OVERVIEW OF RESEARCH ACTIVITIES

In this first year of a three year renewal, substantial progress was made on every major item in the renewal proposal. The most obvious facets of this interdisciplinary work on computers and chemistry are research, engineering and applications. On the research side, the computer programs have grown in both chemical and computer science sophistication. On the engineering side, the programs have been made faster and easier to use. On the applications side, the programs have been used by chemists working on biomedical problems at Stanford and elsewhere as aids in their own research (see ref 4).

STRUCTURE ELUCIDATION PROGRAMS

Stereochemistry in CONGEN

The set of computer programs developed at Stanford as tools for molecular structure elucidation are being enhanced by the addition of 3-dimensional structural information. The programs can now deal with some basic geometrical properties of molecules that are essential for understanding their biological significance. Research progress this year has resulted in extensions that allow computation of stereoisomers (alternative structures differing in 3 dimensions but having identical connections among atoms). Thus geometrical variations on structural hypotheses can be presented as well as topological variations.

The proposed first stage in this effort was to write a program which was capable of recognizing the configurational stereochemical features of a molecule and generate all the possible stereoisomers based on these features. This program has been written and interfaced to an experimental version of CONGEN, and is described in detail below. The proposed second stage in this effort is to modify this program to permit generation of stereoisomers which satisfy certain constraints, much as the existing CONGEN program constrains the generation of topological isomers. This ongoing effort is discussed in the section on future plans.

Example

The structure is 3-6-dimethyl-4-octene, a simple hydrocarbon which exhibits double bond and configuration stereochemistry. It has only a small number of stereoisomers because the symmetry of the molecule reduces the total number that are possible.

```
3-6-dimethyl-4-octene

              4
              |
          1-2-3-5=6-7-9-10
                    |
                    8


THERE ARE 6 STEREOISOMERS

          0 1 1 -1
          1 0 1 1
          2 0 1 1
          4 1 1 -1
          5 0 1 1
          6 0 1 1
```

The first number on each row is the canonical label for each stereoisomer. The correspondence is:

0	R-S-trans
1	S-S-trans
2	R-R-trans
4	R-S-cis
5	S-S-cis
6	R-R-cis

The second number on each row tells whether this particular stereoisomer is achiral (1) or has an enantiomer (0). Enantiomeric pairs are listed on consecutive rows. The final two numbers on each row indicate the symmetry group of each stereoisomer. Those with 1 1 have rotational symmetry and those with 1 -1 have a plane of symmetry.

Future Plans

The following features (at least) will be added to the existing program:

1) Designations of stereocenters as either R or S based on constitutional priorities only. This will be for aid in interpretation only as these designations are not useful internally to the program.

2) Recognition of cis and trans double bonds for the same reason.

3) Stereoisomer output which is interpretable and compatible with character terminal output. This will most likely be done in conjunction with the existing drawing program. The compatibility with character based terminals is a strength of CONGEN at present.

4) Versatility in the handling of the stereochemistry of atoms other than carbon. In particular there should be a choice as to whether a nitrogen atom is thought to be able to invert freely.

The second stage of the development in this effort is to give CONGEN the ability to constrain stereoisomer generation. The algorithm of the generator was designed so that a number of useful constraints, particularly concerning relative stereochemistry between stereocenters can be applied prospectively. That is, the undesired stereoisomers would not be generated. Other constraints, such as those which involve the symmetry of the stereoisomers can be applied during the generation. Finally, there will certainly be some constraints which have to be applied after generation.

Constraints Interpretation

The area of automatic interpretation of constraints in CONGEN structure elucidation problems is interesting and important for two reasons: 1) we want to free the chemist as much as possible from having to understand CONGEN's method of building structures; and 2) problems can be solved much more efficiently if CONGEN can perform some preliminary examination of them and find an alternative, efficient way to solve the problem. Our first efforts in this direction have resulted in what we call the "GOODLIST interpreter", which is designed to make more efficient use of information about required (GOODLIST items plus Superatoms) structural features of an unknown molecule.

Introduction to Method

It is characteristic of structure elucidation based on data from physical and chemical methods that much structural information is redundant. Physical methods, for example, are frequently complementary. One technique provides structural information which can be used to elaborate information gathered by another. The collection of partial structures present in an unknown derived by such methods frequently contain atoms or groups of atoms shared among two or more partial structures. Chemists must take this into account when considering how the partial structures might fit together to yield the structure of an unknown compound. We are developing a method which is designed to demonstrate the task of determining potential overlaps in partial structures.

Stated in the simplest terms, the method should translate the constraints on desired structural features, or GOODLIST constraints, into new sets of partial structures which incorporate the features at the beginning of the structure generation procedure.

Future Directions

Initially, the GOODLIST substructures specified as constraints will be incorporated automatically at the beginning of the problem as described above. Within a short time, the method of specification of a problem will be changed to include only the empirical formula together with inferred partial structures without regard to overlaps, leaving to the program the task of determining those overlaps and specifying the set of problems to solve.

Incorporation of BADLIST (undesired structural features) substructures in the procedure is a necessary next step. Subsequently we will attack the problem of discerning constraints which are implied by the input data, including detection of unclear or ambiguous statements about a structure. The constraints interpreter should be capable of a dialog with the chemist using CONGEN to clarify such points prior to structure generation.

Experiment Planning Program

Now that CONGEN gives us the capability of constructing all plausible candidates under an initial set of constraints, the next problem is to provide the chemist with some assistance in rejecting incorrect candidates and focussing on the correct structure. This process must involve the examination of the candidates to determine their common and unique features, and the designing of experiments to differentiate among them.

The initial work on this problem has begun by providing a new function, the EXAMINE function, which gives a chemist the ability to survey sets of structures for particular combinations of substructures, ring-systems etc. This function has now been incorporated into the CONGEN program. EXAMINE allows structures to be segregated on the basis of combinations of (desired or undesired) structural features. For example, EXAMINE can be used to segregate structures which possess feature A or B, or generally, any arbitrary Boolean expression of relationships among structural features.

More elaborate functions for automatically identifying discriminating features in sets of structures are being developed. Currently, these experimental routines (contained within the "PLAN" program) can be used to analyze functionality, or to identify differences in the ways that superatoms have been imbedded in structures. These routines will shortly be capable of exploiting a simplified library of chemical/spectral tests for particular substructural features; this will allow the program to identify possible discriminating experiments.

The Reaction Chemistry Program

During the past year we have made progress in developing the reaction chemistry program, REACT, into a working tool for laboratory chemists.

REACT is designed to carry out representations of chemical reactions on representations of chemical structures. Reactions, defined by the chemist using the program, are carried out in the synthetic direction as opposed to the retrosynthetic direction of programs for computer-aided synthesis. (6) In structure elucidation problems, the set of structures undergoing reaction is the current set of candidate structures for an unknown. It is clear, however, that the program can also be used effectively in following reactions of a single, known compound participating in a complex sequence of reactions. For example, we showed (ref 22) that CONGEN together with REACT provides a convenient method for studying acid catalyzed rearrangements such as the conversion of tetrahydrodicyclopentadiene to adamantane. In that example, the complete set of isomers was generated by CONGEN. Subsequently, a one-step reaction carried out on each isomer afforded the complete rearrangement graph.

During the past year the structure of the program was revised significantly to include commands and internal operations which more closely parallel laboratory procedures. The new version has been described briefly and some applications of REACT to mechanistic problems have been discussed (ref 24).

Whether applied to mechanistic studies or to structure elucidation problems, REACT's soperatins are based on procedures such as (1) carrying out reactions in a variety of ways, (2) separation of products, (3) labelling of "flasks" for products, (4) carrying out further reactions on products, and (5) testing the contents of flasks against structural constraints.

Structural information obtained in the laboratory on the content of a product flask represents constraints not only on the identity of the product, but also on the identity of the precursor and its precursor and so forth throughout an entire reaction sequence. REACT allows structural statements to be made as constraints on the contents of any flask in a reaction tree. The program then determines automatically the structural implications of the constraint throughout the reaction sequence.

Mass Spectral Prediction and Ranking

Predicting Spectra Using MSRANK and the Half-Order Theory

The MSRANK program has been incorporated as part of CONGEN, but is not yet available for general use. During the past year we have been giving the program some extensive tests to determine its scope and limitations. We have studied the following classes of compounds (all closely related to current research problems): 1) marine sterols; 2) substituted pregnanes; 3) aliphatic and aromatic esters; and 4) macrolide antibiotics.

MSRANK is a powerful filter for eliminating from further consideration structures which cannot yield the observed mass spectrum for an unknown by "reasonable" fragmentation pathways. The greater the structural diversity of isomeric candidates for an unknown, the better the performance of MSRANK in focussing in on the correct structure. When the structures are quite similar, for example when they have been constructed from the same set of superatoms and few remaining atoms, the ranking by MSRANK is quite similar (as one might expect). When this situation occurs, the chemist must still consider the top 10 - 50 percent of the structures as possibilities, depending on the distribution of scores.

---

(6) E.J. Corey and W.T. Wipke, Science 166, 178 (1969).

We have added an explanation feature to MSRANK. Upon request the program prints a list of peaks in the observed spectrum which have different "reasonable" explanations. for different candidate structures. Based on this information the chemist can accept the ranking or change the parameters which define his theory of fragmentation to obtain a different ranking. This procedure helps detect and reduce the plausibility of "nonsense" fragmentation processes.

Prediction Using Fragmentation Rules Supplied by Chemists

When the candidate structure is known to belong to a previously investigated class of compounds, then we can use additional information to predict a more precise mass spectrum. This information is in the form of specific fragmentation rules. These rules are described by a subgraph, a break (or cleavage) and related hydrogen or neutral transfers, intensity ranges associated with rules and a parameter describing the confidence in a rule. We are working on a program which allows the user to enter rules defining his theory of mass spectral fragmentation.

The next step is to explore ways to compare a predicted and an observed spectrum. We are experimenting with different ranking functions (see section 4) and developing a program which will allow the user to define in a simple mathematical equation his individual ranking function. The problem of ranking candidate structures based on spectrum comparison is closely related to the problem of library search. In our case, however, we do not have authentic spectra of our structural candidates in most instances. The density of a predicted spectrum for a candidate is quite low because we do not attempt to predict the complete spectrum. Rather, we predict major fragmentations. This fact must be taken into account in designing a function to rank candidates based on comparison of their predicted spectra to the spectrum of the unknown.

Molecular Ion Determination

The original MOLION program (7) was based upon the postulate: "There exists at least one SECONDARY LOSS in a spectrum that will match a PRIMARY LOSS from the molecular ion irrespective of whether the molecular ion is present in the spectrum."

Given this postulate, then one method of generating candidate masses for a molecular ion (M+) is to identify all possible secondary losses apparent in a spectrum, and then to add each of these losses to the masses of those ions observed in the high mass region of the spectrum. This, together with some refinements, was the basis of the original MOLION program.

There are, however, a number of problems with the algorithm used in MOLION. The most crucial problem is that the algorithm requires good spectra! Impurities such as column bleed or co-eluting minor components can result in ions that would

---

[7] R.G.Dromey, B.G.Buchanan, D.H.Smith, J.Lederberg and C.Djerassi. "Applications of Artificial Intelligence to Chemical Inference. XIV. A General Method for Predicting Molecular Ions in Mass Spectra." Journal of Organic Chemistry, 40, 770 (1975).

constitute bad losses --- causing the rejection of distinct and well supported molecular ions recorded in the spectrum. Further, the program did not allow the user to modify the "bad loss" set, nor to have access to the molecular ion scoring mechanisms. These scoring mechanisms incorporated a considerable measure of class dependency. Thus when testing a candidate M+, the program could modify the score associated with the M+ by the intensity combination formula: e.g. a mass difference of 101amu between the candidate M+ and an observed ion resulted in a 1.8 times increase in that M+'s score whereas a mass difference of 2 or 16 reduced the score by 85 per cent and a difference of 44, 56, 60 or 72 reduced the score by 25 per cent.

In devising the new version of the molecular ion program, an attempt has been made to recognize and overcome some of these problems. We have made modifications suggested by the above considerations. The resulting program is quite successful and has been incorporated into our sequence of programs for analysis of combined gas chromatographic/mass spectrometric (GC/MS) data.

CONGEN Improvements

During the past year many improvements have been made in the version of CONGEN available for outside use. These improvements allow the user more flexibility and range in the use of existing commands. Further, some new commands have been created which increase the power and utility of CONGEN. The program has become easier to use and more robust. Finally in almost every subsection of the program the user can inspect the computation as it proceeds. This means that fewer long, wasteful computations will be performed. Some of the major improvements are described briefly below.

Error Detection in Substructure Definition

We provide extensive error checking on structural fragments (substructures) defined by the user of CONGEN. If the chemist does not choose to fix the errors, we warn very clearly that the results will be unpredictable or erroneous. We allow the chemist to go ahead on the philosophy that he may have a perfectly good reason for doing what seems to us to be nonsense. We have concentrated our efforts on mistakes which we have observed when other chemists use CONGEN. Moreover, this error checking will serve to reduce substantially the errors made by chemists using the program.

Depth-First Imbedder

The IMBEDDER program was completely rewritten. Four major improvements were implemented and the efficiency of almost all of the different subsections was improved. First, the method of computation was changed from breadth first (all structures delivered at the same time) to depth first (the structures delivered one at a time as they are created). The chemist can now check the computation as it proceeds by using the cntrl-S and cntrl-I features. Use of either feature to interrupt the computation and examine results often will allow the chemist to see that a certain computation is much larger than he anticipated and to stop it before computer time is wasted.

Second, all the constraints testing during imbedding is now done in the SAIL portion of CONGEN and structures violating the constraints are not returned. Previously all structures were returned to the LISP portion of CONGEN before any constraints checking and subsequent pruning were done. This new approach represents a real gain in efficiency because these programs run much faster in SAIL than they do in LISP.

Third, the canonicalization routines were rewritten. New algorithms were found and the process of assigning a canonical number to a structure is now much less costly in terms of time. Further, two structures which are aromatically equivalent are now given the same key (related to its canonical number). Since many different parts of CONGEN use the canonicalizer this resulted in a gain in efficiency for all of them.

Fourth, a change was made so that any number of superatoms can be imbedded at one time. This means when large numbers of superatoms need to be imbedded the chemist can in one set of commands perform the entire task, rather than the more time consuming approach of one-at-a-time. However, the chemist can still choose for reasons of efficiency to imbed a single superatom when special tests on the environment of that superatom are required. This also provides the opportunity for large, multiple imbeddings to be done in batch mode. (The batch command was rewritten so that it would accept multiple superatoms.) The large batch job is then run after midnight when the load average is low.

EDITSTRUC Changes

The RENUMBER command was added to give the chemist flexibility in choosing schemes of numbering the atoms in the structure. There have been internal changes made to the editstruc commands BRANCH, LINK, CHAIN, and DELATS. All involve the method of numbering atoms. It is now possible to create a substructure which has "gaps"(missing numbers) in its numbering to atoms. These changes necessitated some further changes in the routines which prepare and send structures to the IMBEDDER in CONGEN.

BATCH

The BATCH command was rewritten to take advantage of the fact that the new imbedder can accept any number of superatoms to be imbedded. As the system load continues to increase BATCH will become a more attractive option.

CONGEN Reprogramming

We have been investigating the reprogramming of CONGEN into an Algol-like language. The goals of reprogramming are threefold: first, to unify the program into a single language which can be used on a variety of computer systems; second, to begin to compact the program into a manageable, cost-effective size for current time-sharing systems; and third, to improve typical runtimes for CONGEN so that it becomes a more attractive means for scientists to solve structure elucidation problems. A version of CONGEN which fulfills these goals would be useful on a variety of computer systems and could be exported to many different chemical and biochemical laboratories.

THEORY FORMATION PROGRAMS - Meta-DENDRAL

Incremental Learning

In order to allow applying the Meta-DENDRAL program (ref 3) to a wider range of chemically interesting problems, we have begun to remove one of the most important current program limitations - its inability to add piecewise to what it has learned. Meta-DENDRAL must currently process all training data at once, producing a set of rules which cover that data.

Since the amount of training data processed strongly influences the reliability of the learned rules, training on arbitrarily large data sets will allow Meta-DENDRAL to form more accurate rule sets than currently feasible.

Modifying Existing Rules

Rules must be modified so that they become consistent with the new data while remaining consistent with previous data as well. In short, the method involves storing along with each rule a summary of alternate acceptable versions of the rule (those with the same evidential support in the observed training data). The summary of all acceptable versions of a given rule, referred to as the <u>version space</u> (ref 14) of the rule, is useful for a number of tasks associated with rule learning, including incremental learning.

Version spaces provide an explicit representation of the space of all alternate versions of a given rule - i.e. those which cannot be disambiguated by the currently observed training data. As such, version spaces will allow Meta-DENDRAL to reason more thoroughly with the choice among alternate rule versions.

Current Status and Future Work

The incremental learning ability for Meta-DENDRAL is almost fully implemented, but as yet remains untested. Routines for defining and modifying rule version spaces are implemented, as well as the ability to filter out training data explained by a rule set. The major unimplemented portion of the incremental learning scheme is the process for merging new rules into the evolving rule set. The chief issue here is deciding when and how to chose among or merge new rules which are similar to existing rules. We expect to complete implementation and initial testing of the incremental learning ability during 1978.

Among issues associated with the version space approach which we expect to explore during the current grant period are the following:

1) Intelligent selection of new training data from examination of partial results.

2) Applying chemical plausibility information to select a "best" rule version from among those contained in the version space.

3) The extension of current methods for dealing more completely with noisy and ambiguous training data.

4) The use of version spaces for merging similar rules.

New Capability To Emphasize Discriminatory Power

One important intended use of rules formed by Meta-DENDRAL is the prediction of mass spectra for use in structure elucidation: Predicted spectra for a set of candidate structures are compared by computer with the mass spectrum observed for an unknown compound, and on this basis the candidates are ranked according to plausibility. The ability of rules, in this context, to differentiate correctly among candidate hypotheses is called their "discriminatory power." Since the selection criteria previously used by Meta-DENDRAL during the various stages of rule formation did not necessarily correlate with high discriminatory power, it was decided to provide the program with the option of directly emphasizing discriminatory power during rule formation, in order to maximize the usefulness of the resulting rules for purposes of structure elucidation.

This addition to Meta-DENDRAL has now been designed and implemented. The general method employed by the the new option is as follows. Observed mass spectra of the training molecules are analyzed prior to rule generation to determine how diagnostic the various observed peaks are, within the training set, of the molecules that show them. This information is then used during rule formation to compute a measure of discriminatory power for emerging rules. This measure is used, in combination with other criteria, to guide the search during rule generation, and to control the modification and selection of rules during the later phases of processing.

Preliminary testing of this new rule formation scheme on the monoketoandrostanes produced rules of considerably greater discriminatory power within that family than had been produced in earlier work with Meta-DENDRAL, even though the training set used was only half as large as that used earlier. This "discrimination option", now integrated with the new template-processing capability, is currently being further tested on a group of aromatic esters to determine whether the rules formed are consistent with what is known about the fragmentation modes of those molecules, and whether the rules have significant discriminatory power outside the training set used to form them.

Data Selection Program

Good inductive generalizations depend on variety in the data set. This is no less true in the context of rule formation by Meta-DENDRAL. Whether the goal is to discover rules of high generality or high discriminatory power, one's chances of achieving this goal appear to increase with increasing variety of training instances. This suggests that it would be useful to have a data selection program that would select the subset of the potential training molecules which has the greatest variety, in some appropriate and well-defined sense. A preliminary version of such a program has been implemented, and experiments with it will soon be underway.

Feedback Loops

The RULEGEN program is capable of accepting previously defined rules as a means of filtering the evidence obtained from INTSUM before the evidence is used for rule formation. As well as providing a convenient and natural feedback mechanism for the program, this facility also allows rules obtained from other sources to be used to reduce the space which the program must examine to find rules for a given set of data. In this manner, the program is able to focus attention on evidence which is not already explained by any of the rules which it is given. Tests are in progress to determine the limitations of this approach.

Stability Rules in INTSUM and RULEGEN

The programs have been generalized to allow the analysis of the mass spectral data from the point of view of determining rules about stable bonds, i.e., lack of fragmentation in a molecule as well as fragmentation. Just as peaks are evidence of fragmentation in a structure, absence of peaks is evidence that certain fragmentations have not occurred.

The programs are now capable of examining the original data from either point of view and proposing rules of behavior of the molecules from that point of view. Further work remains to be done to carry this generality through the processing performed in RULEMOD and then in conducting experiments to determine the usefulness of stability analysis.

Carbon-13 Work

The work described in this section was accomplished in conjunction with work on structure elucidation and theory formation programs (sections 2 and 4).

Carbon-13 nuclear magnetic resonance (CMR) has developed into an important tool for the structural chemist. A natural abundance CMR spectrum which is fully proton decoupled consists of a number of sharp peaks which correspond to the resonance frequencies in an applied magnetic field of the various types of carbon atoms present. A C-13 shift is the amount an observed peak is shifted from that of a reference peak, usually tetramethylsilane (TMS).

During the past year we continued work on an extension of Meta-DENDRAL which allows the program to form rules in the domain of CMR spectroscopy. We also wrote a second program which applies CMR rules to structure elucidation problems. Rules generated from a combined set of paraffins and acyclic amines have been used to successfully identify the C-13 NMR spectra of molecules not in the training set data. The introduction of a limited set of stereochemical terms to the rule generation procedure demonstrated the feasibility of extending the method to more complicated systems. A description of the rule formation and structure elucidation programs is given in (ref 16). Results are presented there for the combined set of paraffin and acyclic amines, as well as for a combined set of trans decalins and monohydroxylated androstanes.

Molecular structure elucidation is accomplished by our program by selecting a shift (peak) in the observed spectrum, then finding the rules which are possible explanations for this shift. The rules selected postulate partial substructures which might be in the molecule. These substructures are then assembled jigsaw puzzle fashion to construct the final molecule. Constraints stemming from both the observed spectrum and information associated with each rule are used to constrain the process so that only "reasonable" structures will be considered.

The structure elucidation program has been run on several test cases using unknown paraffin and acyclic amine spectra with reasonable success. This program is described in detail in (ref 16).

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

Use of CONGEN via SUMEX

The following individuals were among those listed in the DENDRAL annual report as persons who use CONGEN or who have requested information about access.

Dr. Peter Gund of Merck, Sharpe and Dohme Laboratories

Professor Richard E. Moore of the University of Hawaii

Dr. Jean-Claude Braekman of the University of Brussels

Dr. Martin Huber, a postdoctoral fellow in Professor Wipke's SECS group

Professor Kurt Mislow of Princeton University

Professor Weiss, head of the Department of Chemistry at Northeastern University.

Dr. Stan Lang of Lederle Labs' Infectious Disease Research Section

Dr. Leon Goldman and Dr. Babu Venkataraghavan, also at Lederle

Professor E.J. Eisenbraun of Oklahoma State University

Dr. David Pensak of Dupont in Wilmington, Delaware

Dr. Milton Levenberg of Abbott Laboratories

Kent Morrill of Tennessee Eastman

Dr. Gretchen Schwenzer of Monsanto

Dr. Robert Shapiro of New York University

Dr. Henry Stoklosa of Ciba-Geigy

Dr. Geza Szonyi of Polaroid Corporation

Drs. D. Williams and R. McGrew from Dow Chemical

Interactions with Other SUMEX-AIM Projects

The community of scientists is a valuable resource in itself. We have shared many ideas about both programming and chemistry with several other SUMEX projects. In particular:

> Prof. Todd Wipke of the Chemical Synthesis Project has spent part of his sabbatical quarter at Stanford (in a DENDRAL office) and has attended several of our group meetings. We have interacted with other members of his group on common problems of graph-theory as it relates to representation and manipulation of chemical structural information.
>
> Dr. Robert Shapiro spent a week visiting us to learn about CONGEN and related programs and their application to structures of unknown compounds obtained from reactions of various compounds with DNA nucleosides.
>
> Bill White, a senior programmer for DENDRAL, has been working with Penny Nii, Bill vanMelle and others on programmers' aids for managing large INTERLISP programs.
>
> The MOLGEN project has as its primary goal assisting geneticists in experiment planning. One of the new research areas of our project (see above) is also experiment planning, in our case to assist chemists in solving new structures. We have interacted with the MOLGEN project regularly to exchange ideas and discuss methods.
>
> Although Ray Carhart has been in Edinburgh this year, we have continued close interaction with him because he is able to send and receive messages from us and run programs on SUMEX.

Critique of Resource Management

DENDRAL's View of the SUMEX Resource

Our research efforts depend heavily on adequate computer resources in order that both program development and application of the program to structural problems can be carried out. The interactive nature of program development and application demands a time-sharing environment where such interactions are possible. The SUMEX resource has been an ideal vehicle, not only in its support of time-sharing but also in the variety of languages, editors and system functions SUMEX provides which can be employed to solve special problems.

We are currently approaching the limits of our share of the SUMEX resource. The combination of new developments and our desire to offer at least trial access to a broad community of collaborators comes at a time when the demands on the resource from all projects are taxing its capabilities. We have alleviated this problem somewhat by voluntary scheduling of our development efforts, including shifting some work away from prime time. We also encourage our collaborators to avoid prime time use whenever possible, and have provided some batch capabilities to run long computations overnight. However, despite these efforts, there is no question in our minds that the pace of new developments and applications is slowed somewhat by the demands on SUMEX. We welcome any augmentation of the resource, for example, a smaller machine for applications packages, which would alleviate the current situation.

III. RESEARCH PLANS

Long Range Goals

We are developing an integrated package to aid scientists in the elucidation of molecular structure of compounds of biomedical significance. Specific steps within the current 3-year grant period are detailed in the research summary sections above.

Justification and Requirements for Continued SUMEX Use

We have successfully demonstrated to the NIH the biomedical relevance and importance of the research. The last site-visiting team was composed of about a dozen nationally known chemists and computer scientists, who were extremely thorough in their examination of the goals and methods of the project. Their endorsement is taken as strong support for the biomedical relevance of our work as well as support for the technical details of our methods.

The DENDRAL programs have been developed in INTERLISP because of the extreme flexibility of the language and the programming aids it offers. We will continue to exploit these features for trying new ideas and developing first versions of new programs, even as we map the better developed parts of our program into a more compact (and more rigid) language.

Use of Other Computational Resources

As mentioned in the research summary, we are developing a version of CONGEN for the NIH/DCRT machine and for the NIH/EPA Chemical Information System computer. We have arranged for time on these machines to test the programs. These arrangements will not affect our use of SUMEX but will ease the outside demand for access to SUMEX by chemists wanting to run CONGEN.

Recommendations for Resource Development

see Critique of Resource Management, above.

IV. FUNDING

National Institutes of Health Biotechnology Resources
Program Grant RR-00612

Principal Investigator: Carl Djerassi, Principal Investigator
Professor of Chemistry, Stanford University

Budget for Year 1 from 5/1/77 through 4/30/78
Total Direct Costs: $218,580.

3-Year Summary from 5/1/77 through 4/30/80
3-Year Total Direct Costs $698,399.

REFERENCES

[1] Bruce G. Buchanan and Dennis H. Smith, "Computer Assisted Chemical Reasoning," in E.V. Ludena, N.H. Sabelli and A.C. Wahl (eds.), Computers in Chemical Education and Research, New York: Plenum Press, 1977. P. 401

[2] Bruce G. Buchanan, "Issues of Representation in Conveying the Scope and Limitations of Intelligent Assistant Programs," in D. Michie (ed.), Machine Intelligence 9, forthcoming.

[3] Bruce G. Buchanan and Tom Mitchell. "Model-Directed Learning of Production Rules," in D.A. Waterman and F. Hayes-Roth (eds.), Pattern-Directed Inference Systems, New York: Academic Press, forthcoming.

[4] Bruce G. Buchanan and Edward A. Feigenbaum, "DENDRAL and Meta-DENDRAL: Their Applications Dimension," Artificial Intelligence, forthcoming.

[5] Raymond E. Carhart and Dennis H. Smith, "Applications of Artificial Intelligence for Chemical Inference XX. Intelligent Use of Constraints in Computer-Assisted Structure Elucidation", Computers and Chemistry, 1, 79 (1976).

[6] Raymond E. Carhart, "A Model-Based Approach to the Teletype Printing of Chemical Structures," Journal of Chemical Information and Computer Sciences, 16, 82, 1976.

[7] C.J. Cheer, D.H. Smith and C. Djerassi, and B. Tursch, J.C. Braekman and D. Daloze, "Applications of Artificial Intelligence for Chemical Inference XXI: The Computer-Assisted Identification of [+]Palustrol in the Marine Organism Cespitularia sp., aff. Subviridis", Tetrahedron, 32, 1807 (1976).

[8] C. Djerassi, R. M. K. Carlson, S. Popov and T. H. Varkony. Sterols from Marine Sources. In press.

[9] R. G. Dromey, Mark J. Stefik, Thomas C. Rindfleisch, and Alan M. Duffield, "Extraction of Mass Spectra Free of Background and Neighboring Component Contributions from Gas Chromatography/Mass Spectrometry Data," Analytical Chemistry, 48, 1368, August 1976.

[10] S. Popov, R. M. K. Carlson, A-M. Wegmann and C. Djerassi. Occurrence of 19-Nor Cholesterol and Homologs in Marine Animals. Tetrahedron Lett., 3491 (1976).

[11] S. Popov, R. M. K. Carlson, A-M. Wegmann and C. Djerassi. Minor and Trace Sterols in Marine Invertebrates. 1. General Methods of Analysis. Steroids, 28, 699 (1976).

[12] Gretchen M. Schwenzer, "Applications of Artificial Intelligence for Chemical Inference. XXVI Analysis of C-13 NMR for Mono-Hydroxy Steroids

Incorporating Geometric Distortions," Journal of Organic Chemistry, forthcoming.

[13] T.M. Mitchell and G.M. Schwenzer, "Applications of Artificial Intelligence for Chemical Inference XXV. A Computer Program for Automated Empirical 13C NMR Rule Formation," Organic Magnetic Resonance, forthcoming.

[14] Tom M. Mitchell, "Version Spaces: A Candidate Elimination Approach To Rule Learning," Proceedings of the Fifth IJCAI, 1, 305, August 1977.

[15] James G. Nourse, "Generalized Stereoisomerization Modes," Journal of the American Chemical Society, 99, 2063, 1977.

[16] Gretchen M. Schwenzer and Tom M. Mitchell, "Computer Assisted Structure Elucidation Using Automatically Acquired 13C NMR Rules," in D. Smith, (ed.), Computer Assisted Structure Elucidation, ACS Symposium Series, Vol. 54:58, 1977.

[17] D.H. Smith, (ed.), "Computer-Assisted Structure Elucidation." (ACS Symposium Series 54). Washington, D.C.: American Chemical Society, 1977.

[18] R.E. Carhart, T.H. Varkony, and D.H. Smith, "Computer Assistance for the Structural Chemist," in "Computer-Assisted Structure Elucidation," D.H. Smith, (ed.), American Chemical Society, Washington, D.C., 1977, p. 126.

[19] D.H. Smith, M. Achenbach, W.J. Yeager, P.J. Anderson, W.L. Fitch, and T.C. Rindfleisch, "Quantitative Comparison of Combined Gas Chromatographic/Mass Spectrometric Profiles of Complex Mixtures," Anal. Chem., 49, 1623 (1977).

[20] D.H. Smith and P.C. Jurs, "Prediction of 13C NMR Chemical Shifts," J. Am. Chem. Soc., submitted for publication.

[21] Dennis H. Smith and Raymond E. Carhart, "Structure Elucidation Based on Computer Analysis of High and Low Resolution Mass Spectral Data," in M.L. Gross (ed.), Proceedings of the Symposium on Chemical Applications of High Performance Spectrometry, Washington, D.C.: American Chemical Society, in press.

[22] T.H. Varkony, R.E. Carhart, and D.H. Smith, "Applications of Artificial Intelligence for Chemical Inference XXIII. Computer-Assisted Structure Elucidation. Modelling Chemical Reaction Sequences Used in Molecular Structure Problems," in W.T. Wipke, (ed.), Computer-Assisted Organic Synthesis, Washington, D.C.: American Chemical Society, 1977.

[23] Tomas H. Varkony, Raymond E. Carhart, and Dennis H. Smith, "Computer Assisted Structure Elucidation, Ranking of Candidate Structures, Based on Comparison Between Predicted and Observed Mass Spectra," in Proceedings of the Twenty-Fifth Annual Conference on Mass Spectrometry and Allied Topics, Washington, D.C., 1977.

[24] Tomas Varkony, Dennis Smith, and Carl Djerassi, "Computer-Assisted Structure Manipulation: Studies in the Biosynthesis of Natural Products," Tetrahedron, forthcoming.

[25] T.H. Varkony, R.E. Carhart, D.H. Smith, and C. Djerassi, "Computer-Assisted Simulation of Chemical Reaction Sequences. Applications to Problems of Structure Elucidation," J. Am. Chem. Soc., submitted for publication.

[26] Annemarie Wegmann, "Variations in Mass Spectral Fragmentation Produced by Active Sites in a Mass Spectrometer Source," Analytical Chemistry, forthcoming.

4.2.3 AGE PROJECT

AGE ("Attempt to Generalize")

H. Penny Nii
Edward A. Feigenbaum
Computer Science Department
Stanford University

Isolate inference, control and representation techniques from previous knowledge-based programs; reprogram them for domain independence; write a rule-based interface that will help a user understand what the package offers and how to use the modules; and make the package available to other members of the AIM community and labs doing knowledge-based systems development, and the general scientific community.

I. SUMMARY OF RESEARCH PLAN

Technical Goals

The goal of this new effort is to construct a computer program to facilitate the building of knowledge-based systems. The design and implementation of the program will be based primarily on the experiences gained in building knowledge-based systems at the Heuristic Programming Project in the last decade. The programs that have been, or are being, built are: DENDRAL, meta-DENDRAL, MYCIN, HASP, AM, and MOLGEN and CRYSALIS. Initially, the AGE program will embody methods used in these programs. However, the long-range objective is to integrate methods and techniques developed at other AI laboratories. The final product is to be a collection of building-block programs combined with a knowledge-based front-end that will assist the user in constructing knowledge-based programs. It is hoped that AGE can speed up the process of building knowledge-based programs and facilitate the dissemination of AI techniques by: (1) packaging common AI software tools so that they do not need to be reprogrammed for every problem; and (2) helping people who are not knowledge-engineering specialists to write knowledge-based programs.

Medical Relevance and Collaboration

The DENDRAL, meta-DENDRAL, MYCIN, MOLGEN, and CRYSALIS, projects have been creating intelligent agents to assist human problem solving in tasks of significance to medicine and biology (see separate sections for discussions of work and relevance). Without exception the programs were handcrafted. This process often takes many years, both for the AI scientists and for the experts in the field of collaboration.

The time has come for automating some of the activities involved in building knowledge-based programs. Close collaboration of domain experts is still necessary to provide the knowledge base, but we can reduce the system building and programming time of the AI scientists. Since we view our science as an empirical science, in which many programming experiments are conducted before we produce programs suitable for a task, reducing the programming and experimenting time would significantly reduce the time required to build knowledge-based systems.

Progress Summary

AGE SYSTEM ORGANIZATION

AGE itself is a knowledge-based system organized around five subsystems, each containing its own distinct knowledge. Most of the knowledge is about itself - what facilities it has, and how to use those facilities. Figure 1 shows the general interrelationship among these subsystems. When the system is completed,it is intended to be self-contained.

```
  SYSTEM KNOWLEDGE      SUBSYSTEMS        RESULT
                            |
   ___________         _____V______
  | MANUAL    |....>| TUTOR       |
  |___________|...  |_____|______|
                 .        |
   ___________   .     _____V______     ____________
  | DESIGN    |....>| DESIGN      |....>|USER SYSTEM |
  | SCHEMA    |...  |             |    |  DESIGN    |
  |___________| .   |_____|______|    |______↑_____|
                 .        |<..................
   ___________   .     _____V______     ____________
  | BUILDING  |....>| ACQUISITION|....>|  USER      |
  | BLOCKS    |     |             |    | SYSTEM     |
  |___________|     |____________|    |______|_____|
                          |<.................
                     ______V_____           .
                    |  TEST      |          .
                    |______|_____|          .
                          |<.................
   ___________       ______V_____
  |EXPLANATION|....>|EXPLANATION |
  | SCHEMA    |     |            |
  |___________|     |____________|
```

Figure 1 AGE System Organization

The five subsystems and their current status are described below:

1. TUTOR:

Description: TUTOR's function is to guide the user in browsing through its knowledge base called the MANUAL. The MANUAL contains a general description of the building-block components on the conceptual level, a more specific description of the implementation of these concepts within AGE, description of how these components are used and how they can be constructed by the user, and various examples.

Status: The prototype TUTOR system has been written. The MANUAL is approximately 50% complete for the current version of AGE.

2. DESIGN:

Description: The function of the DESIGN subsystem is to guide the user in designing and constructing his program (henceforth called the "object program". The necessary knowledge is represented in the DESIGN-SCHEMA. Using this schema, the DESIGN subsystem guides the user from one design decision point to another. At each decision point, the user has access to the MANUAL and to advice regarding design decision at that point. This subsystem will also keep track of what has, and what has not been specified. An appropriate ACQUISITION module can be invoked form the DESIGN subsystem so that general design and implementation specifications can be accomplished simultaneously.

Status: The DESIGN-SCHEMA is complete for the currently available building blocks. A prototype DESIGN subsystem has been written to help the user design the object program.

3. ACQUISITION:

Description: For each system component that the user must specify, there is a corresponding acquisition module that will ask the user for task specific information. The acquisition subsystem is guided by DESIGN-SCHEMA as well as more detailed knowledge about the components.

Status: About 80% of the Acquisition subsystem is complete.

4. TEST:

Description: This subsystem contains the testing and debugging facility and will rely heavily on the excellent INTERLISP debugging facility. It will be augmented by traces of the reasoning steps of the object program, and of AGE itself. Eventually the tracing capability will be extended back to the decisions made during the design phase leading to "debugging" on the conceptual level.

Status: Traces of the runs of the object program is available. However, the testing facility is minimal at the present.

5. EXPLANATION:

Description: AGE has enough information for a replay of its execution steps, and it has reasonable justifications for the actions within the building blocks. However, AGE is totally ignorant of the object task domain and has no way to conduct a dialogue about that task. In the future we hope to represent different kinds of explanation that will interface to the knowledge base of the user's domain.

Status: No work has begun on this subsystem.

BUILDING BLOCKS

Although the building-block components have roots in previous programs, they have been carefully selected and modularly programmed to be useable in new combinations. The current AGE system aims to provide the user with a subset of problem solving methods applicable to hypothesis formation [Nii:77, Engelmore:77]. This hypothesis formation framework is based on the HEARSAY MODEL [Erman:75, Lesser:77] and uses the concepts of a a"blackboard" (a globally accessible data structure) and independent sources of knowledge which cooperate to form hypotheses (see figure 2).

Within this framework, currently there are enough components to build object programs for tasks characterized by the need for:

1. integration of multiple, independent sources of knowledge,

2. multi-level representation of solution hypothesis (hierarchically organized hypothesis structure),

3. representation and manipulation of uncertain knowledge using weights assigned to inferences generated by rules,

4. problem-solving methods using goals, expectations, and events,

5. independent focus-of-attention mechanisms.

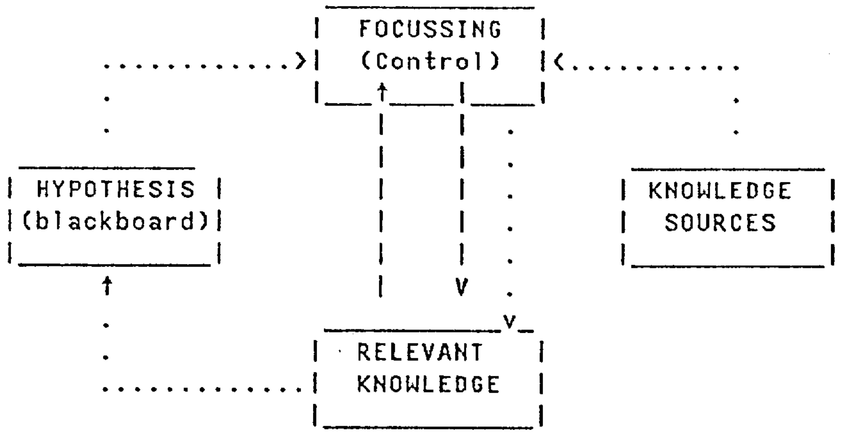

Figure 2. Control flow of the object program - HEARSAY model

APPLICATION OF AGE

Using the current version of AGE (AGE-0), two knowledge-based program are being developed. They are:

1. AGEPUFF: a program to diagnose pulmonary function disorder. One version with approximately 50 rules has been completed. The rules used in this application are those developed for the PUFF project described elsewhere. These rules were modified to be consistent with AGE representation. The current method applied to PUFF rules is and event-oriented invocation of rules. An expectation-oriented rule processing method and MYCIN-like goal-oriented rule processing method are now being applied to the same set of rules.

2. CRYPTO: a program to solve cryptogram puzzles (a "breadboard" domain to assist program development and debugging).

These knowledge-based programs use the currently implemented problem solving framework (collection of building blocks) that was described above. That part of AGE designed to help the user construct his knowledge-based program was used to transform the PUFF rules into AGE acceptable form.

Funding Support Status

The AGE project is currently partially supported under the Heuristic Programming Project contract with the Advance Research Projects Agency of the DOD, contract number MDA 903-77-C-0322, E. A. Feigenbaum, Principle Investigator. During the next year, SUMEX core research funds will provide partial personnel support for approximately 0.9 FTE.

II. RESEARCH PLAN

Research Topics

Two specific long range research activities of the AGE effort are:

1. The isolation of techniques used in knowledge-based systems. It has always been difficult to determine if a particular problem-solving method used in a knowledge-based program is "special" to a particular domain or whether it generalizes easily to other domains. In the currently existing knowledge-based programs the domain-specific knowledge and the manipulation of such knowledge using AI techniques are often so closely coupled that it is difficult to make use of the programs for other domains. Our goal is to isolate the AI techniques that are general to determine precisely the conditions for their use.

2. Guiding users in the initial application of these techniques. Once the various techniques are isolated and programmed for use, an "intelligent front end" is needed to guide users in the application of these techniques. Initially, we assume that the user understands AI techniques and knows what he wants to do, but that he does not understand how to use the AGE program to accomplish his task. A longer-range interest involves helping the user determine what techniques are applicable to his task. That is, we assume that the user does not understand the necessary techniques of writing knowledge-based programs. Some questions to be posed are: What are the criteria for determining if a particular application is suited to a particular problem-solving framework? How does one decide the best way to represent knowledge for a given problem? There are some smaller, but by no means trivial, questions which also need answering. Is there a "best way" to write production rules which would apply to many task domains? Is there a data representation which would cover many tasks? What is the best way to handle differences in the ability of the users of the AGE program?

To correspond to the two general research goals described above, the AGE program will be developed along two separate fronts, both of which are divided into incremental development stages. The first of these fronts is the development of the ability to help build many different types of knowledge-based programs (the "generality" front). The current framework within which object programs can be developed is a variation of the HEARSAY model. The various components of this model are contained in the BUILDING BLOCK/DESIGN/ACQUISITION subsystems. The second front is the development of "intelligence" in the interaction between the user and the AGE program; i.e. moving from dialogues on "how to use the tools in AGE" to "what tools to use" (the "how-to-what" dialogue front). A solution to this problem is reflected in the TUTOR and DESIGN subsystems described above.

Research Plan

The current plan for the development of the AGE program follows:

a. Immediately, augment the current capability of AGE to build HEARSAY-like programs with a capability to build MYCIN-like goal-oriented programs. Add to the dialogue capability, an ability to discuss how to chain rules and how to specify the necessary parameters for the context-tree-like structure. b.

Within the next three years, explore other AI methods for addition to the AGE system. Begin to extract from the user some key characteristics of the task, and using that information begin to suggest appropriate knowledge representations and problem-solving techniques for the user's task. This interactive capability will be limited to match the problem-solving methods available in AGE.

c. Test the utility of the AGE system by developing a more complex application program in some task domain.

References

ENGELMORE:77 Engelmore, R.S. and Nii, H.P., "A knowledge-based system for the interpretation of protein x-ray crystallographic data," Heuristic Programming Project Memo HPP-77-2, January, 1977.

ERMAN:75 Erman, L.D. and Lesser, V.R., "A multi-level organization for problem solving using many, diverse, cooperating sources of knowledge," Proc. 4th IJCAI, 1975, pp.483-490.

LESSER:77 Lesser, V.R. and Erman, L.D., "A retrospective view of the HEARSAY-II architecture," Proc. 5th IJCAI, 1977, pp. 790-800.

NII:77 Nii, H. P. and Feigenbaum, E.A., "Rule-based understanding of signals," Conf. on Pattern-Directed Inference Systems, Hawaii, 1977.

4.2.4 HYDROID PROJECT

HYDROID - Studies in Distributed Processing and Problem Solving

Prof. Gio Wiederhold
Computer Science and Electrical Engineering
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

A. TECHNICAL GOALS

The objective of this research is the development of a methodology for the analysis and the implementation of distributed processing. The primary reason for interest in this area is its potential of multi-processors to break through the speed limitation barriers imposed by uniprocessing systems. Other reasons are expectations of reduced cost of computation and increase in reliability. If such a breakthrough can be achieved then the viability of the applications being developed by other projects using the SUMEX-AIM resource will be enhanced.

The rapid development of processor and communications technology has given rise to a large number of proposals for implementations of networks employing multiple processors. The computations to which these distributed systems are to be applied include heuristic decision-making problems, mathematical modelling, data reduction, and database search, as well as general purpose multi-access computing. There is however a lack of an adequate global understanding of the computational trade-offs implied by network architectures.

In order to complement the experimental results of other investigators and broaden their applicability to the system-design decision-making process, we are developing a general framework for the study of processor interaction in parameters from programs which specify the computations, rules to parameterize descriptions of networks of processors, and procedures to calculate expected system performance from these parameter sets. The framework is to be sufficiently powerful so that, when it is validated, the methods will be able to assist in the a priori assessment of the potential performance of new system alternatives or of systems with improved system components.

One of the primary tools we are using to analyze the interaction between computations and distributed processor networks is simulation. The behaviour of processor network nodes, interprocessor control and task flow, and problem decomposition all require simulation at different levels of abstraction. Analytic queuing models may provide insight into relationships in networks, but are not adequate to provide quantitative results. Simulation is not seen as the end product of the study, but as a means to develop and assess the validity of our model of the interaction of computations and processor network architecture. Where possible, mathematical results are used to assess the validity of model simulations. Most actual simulations, since they require extensive computing resources, are being done using other computers at Stanford.